Alignment of Language Agents

Year: 2021 | Citations: 200 | Authors: Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik

---

Abstract

For artificial intelligence to be beneficial to humans the behaviour of AI agents needs to be aligned with what humans want. In this paper we discuss some behavioural issues for language agents, arising from accidental misspecification by the system designer. We highlight some ways that misspecification can occur and discuss some behavioural issues that could arise from misspecification, including deceptive or manipulative language, and review some approaches for avoiding these issues.